The Survival Kit

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

The Survival Kit: Software to analyze survival data including possibly correlated random effects

G. Mészáros[a,*], J. Sölkner[a], V. Ducrocq[b]

[a] Division of Livestock Sciences, University of Natural Resources and Life Sciences, Vienna, Gregor-Mendel-Str. 33, A-1180 Vienna, Austria
[b] INRA, UMR 1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy-en-Josas, France

ARTICLE INFO

Article history:
Received 17 November 2011
Received in revised form
10 January 2013
Accepted 13 January 2013

Keywords:
Survival analysis
Proportional hazards
Frailty model
Correlated random effects

ABSTRACT

The Survival Kit is a Fortran 90 Software intended for survival analysis using proportional hazards models and their extension to frailty models with a single response time. The hazard function is described as the product of a baseline hazard function and a positive (exponential) function of possibly time-dependent fixed and random covariates. Stratified Cox, grouped data and Weibull models can be used. Random effects can be either log-gamma or normally distributed and can account for a pedigree structure. Variance parameters are estimated in a Bayesian context. It is possible to account for the correlated nature of two random effects either by specifying a known correlation coefficient or estimating it from the data. An R interface of the Survival Kit provides a user friendly way to run the software.

1. Introduction

The most popular class of survival models is the class of proportional hazard models [1,2], where the hazard of an individual at time t is described as the product of the baseline function and of a positive term which is an exponential function of a vector of covariates $\mathbf{w}'$ multiplied by vector of regression parameters $\boldsymbol{\theta}$. Frailty models are an extension of standard survival analysis models which allows to account for unobserved random heterogeneity [3] or equivalently, to include random effects. These account for an unobserved environmental or genetic effect affecting the hazard of the individual. When two random effects are included (e.g., [4]), these can be independent from each other or related to some degree, leading to the need to estimate correlated random effects. Analyses failing to account for this underlying correlation in survival times are likely to underestimate the variances of parameters [5].

The aim of this paper is to introduce "the Survival Kit", software for survival analysis capable to handle very large amounts of data, with the possibility to account for their right censored or left truncated status in proportional hazards models. The fixed, random and stratification variables can be time dependent. The estimation of variance components is done in a Bayesian framework and is based on a Laplace approximation of the marginal posterior density of these parameters, from which a modal point estimate can be obtained. Various modeling possibilities are shown in [4], including stratified and frailty survival models with simultaneous estimation of variances for two random effects, center and interaction of treatment by center. The first random effect corrected for deviation centers from the overall baseline hazard, while the second was to deal with deviation of each center from the overall treatment effect. When required, the first three moments of this posterior density can be estimated and the full posterior density can be approximately constructed and visualized. The program was originally written in Fortran 90 for computational

* *Corresponding author*. Tel.: +43 1 47654 3259.
E-mail address: gabor.meszaros@boku.ac.at (G. Mészáros).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.01.010

efficiency on very large datasets. An R interface was added to provide easier usage and graphical capabilities.

In Section 2 we present the statistical model. In Section 3, the Survival Kit is described and in Section 4 two illustrative applications are presented, one using real infant mortality data, the other using simulated data, and computing variances and covariance of correlated random effects.

2. Theoretical background and computational methods

This section presents a brief overview of the methods used in the Survival Kit. More detailed information can be found in [4,6].

Proportional hazard models and their extension to include random effects describe the hazard function of each individual $\lambda(t)$ (i.e., its limiting probability of dying at time t, given it is still alive just prior to t) as the product of a baseline hazard function and a positive (exponential) function of explanatory covariates.

The model is specified as:

$$\lambda(t; \mathbf{x}(t), \mathbf{z}(t)) = \lambda_0(t) \exp\{\mathbf{x}(t)'\boldsymbol{\beta} + \mathbf{z}(t)'\mathbf{s}\} \quad (1)$$

where $\boldsymbol{\beta}$ and $\mathbf{s}$ are vectors of fixed regression coefficients and random effects. The second part, $\exp\{\mathbf{x}(t)'\boldsymbol{\beta} + \mathbf{z}(t)'\mathbf{s}\}$, represents a stress-dependent term specific to the animals with fixed covariates $\mathbf{x}$ and random covariates $\mathbf{z}$. Both the fixed and random covariates can be time dependent. Only stepwise functions of time are considered for $\mathbf{x}(t)$ or $\mathbf{z}(t)$, i.e., $\mathbf{x}(t)$ and $\mathbf{z}(t)$ are supposed to remain constant over intervals $[t_i, t_i + 1[$. The first part $\lambda_0(t)$ is the baseline hazard function. It is left unspecified in the Cox model [1] or it can take a parametric form as in Weibull model shown in (2).

$$\lambda_0(t) = \lambda\rho(\lambda t)^{\rho-1} \quad (2)$$

where λ and ρ are the shape and scale parameters of the Weibull distribution [2].

The baseline hazard function can be unique or can differ between groups of individuals. The time scale can be divided into several intervals using stratified models, with specific baseline hazards with a separate origin for each, defining a piecewise (e.g., piecewise Weibull) model. This is useful to evaluate hazards with a repetitive pattern. One example is the modeling of culling in dairy cows which clearly follows a particular within lactation pattern [7].

In case of discrete time scale (i.e. with very few distinct time values), there are often many failures occurring at the same time, leading to "ties" between failure times. In such case, the Cox model is no longer valid: if m failure times are tied at time l and n individuals are at risk just prior to l, the partial likelihood contribution involves a summation over all possible subsets of size m from the n at risk, which makes the choice of a Cox model for the discrete time measures inadequate and computationally demanding. Prentice and Gloeckler [8] proposed another approach, the "grouped data model" based on [9]. They assumed that the actual failure times occur in a number of intervals (e.g., years) $[0 = \tau_0, \tau_1), [\tau_1, \tau_2), \ldots [\tau_{k-1}, \tau_k), \ldots$ and that the risk of failure is constant within each interval. All failures occurring in the same interval $[\tau_{k-1}, \tau_k)$ are "grouped", and the attached failure time is k. They also assumed that censoring occurs at the end of each interval. The estimation procedure they proposed was included in the Survival Kit, using a reparameterization described in [10]. Indeed, it is possible to rewrite the model as an exponential regression model including an additional time-dependent effect changing at the beginning of each new interval (see [10] for details).

Technically, the hyperparameters of the prior distribution of random effects (e.g., genetic variance) are estimated from their marginal posterior density [6]. The latter can be obtained through the exact algebraic integration of the random effect out of the joint posterior density when the random effect is assumed to follow a log-gamma distribution. However this is not possible when a normal (or multivariate normal) distribution is used for random effects, for example genetic effects of related animals. Instead, an approximate integration can be implemented using a Laplace approximation. Then, assuming the hyperparameters known, the estimates of all other parameters are obtained as the mode of their joint posterior density. This maximization is done using a limited memory quasi-Newton approach [11] which only requires the computation of the vector of first derivatives of the function to maximize.

For very large applications and models involving correlated random effects, the quasi-Newton approach may converge very slowly. In this case, a full Newton–Raphson algorithm (using both the first and the second derivatives of the function to maximize) can be used to guarantee convergence in a much smaller number of (computationally more expensive) iterations. Also a combination of both quasi-Newton and full Newton–Raphson algorithms is possible and even advisable when good starting values are not available.

Finally, it is also possible to jointly estimate the variance of two random effects using a derivative free algorithm. A normal distribution can be assumed for each level of both random effects. When individuals are (genetically) related, all relationships can be accounted for, assuming a multivariate normal distribution with a (co)variance matrix proportional to $\mathbf{A}$, their relationship matrix [12]. These random effects can be independent from each other [4], but it is also possible to account for their correlated nature as in [5], for example when they correspond to time-dependent effects, for example two genetic effects influencing differently the trait of interest in early and late life. In this case, the two random effects should have the same number of levels. The variances of the random effects and their correlation coefficient could be specified (in case of availability of good prior estimates) or estimated simultaneously with the program.

3. Computer program

3.1. *General description*

The Survival Kit has been developed since its first release in 1994, gradually adding possibilities of stratification and different model types, notably the possibility to model correlated random effects as its latest feature. It is heavily used mostly in the animal breeding community, demonstrated by over

280 scientific papers using the Survival Kit, as of December 2012. This part will describe the program in its current state (for more information about changes between versions, see [13–15]).

The Survival Kit is written in Fortran 90 and can process very large amounts of data, for example several millions of individual records in national breeding value evaluations in cattle [7]. Computational time depends on the actual size of the dataset, the model complexity, whether or not (correlated) random effects are used and whether hyperparameters are assumed to be known or need to be estimated. The program handles time dependent random effects as well (e.g., two different (possibly correlated) genetic effects affecting the risk of the same individual during two different periods). It runs on every operating system after compilation. The source code is freely available on the website mentioned at the end of this paper.

The Survival Kit handles any number of fixed (possibly time dependent) continuous or discrete covariates with any number of classes. Time dependent covariates are assumed piecewise constant over time intervals. Therefore, records with time dependent effects should include the time of change and the new value of the covariate, as many times as the covariate changes, for example if the pth variable in the input dataset has a value of v_0 at $t = 0$, any change in value is indicated in the input data file as a triplet (p, t_j, v_j), meaning that this p^{th} covariate changes to the new value v_j at time t_j. Any number of triplets can be specified. Consequently, survival records are split into so called "elementary records" internally, each covering only the time span from a change in any covariate to the next. The last column of each record holds the number of triplets for the time dependent variables for that particular record, with 0 if there were no changes for the time dependent effect, or these are not used at all.

It is possible to include random effects into the evaluated model, which can be interaction terms with other covariates and/or time dependent. The program computes a point estimate and its standard error for each level of the random effect. It can also estimate the variance of the random effect as the mode of its posterior density with the possibility to provide also its mean, standard deviation and skewness. The latter three parameters give a more accurate picture about the approximate posterior density of the random effect, which can be visualized by Gram–Charlier approximation [16] as shown in Fig. 5. The knowledge of the mean and standard deviation of this posterior density (or of the whole posterior density) can be used to decide whether the corresponding random effect is statistically different from 0.

There could be any number of random effects in the evaluated model, out of which the variances for at most two could be estimated simultaneously, the others being assumed to be known (to estimate the variance of more than 2 random effects, cyclic maximization can be used). The two estimated variances are assumed to be independent by default, but they can also be correlated. The correlation coefficient can be assumed to be known (e.g., from a previous analysis) or be estimated.

Initially, the Survival Kit had to be first compiled on the local computer and then used specifying the model parameters in a text file. Recently an interface to R [17] has been developed to simplify the usage of the software and provide visualization options. With the R interface, it is possible to create graphs such as: Kaplan–Meier curve with 95% confidence intervals, graphical test of the Weibull hazard assumption (e.g., *log(−log(KM estimate))* vs. *log(time)*), survival function and cumulative hazard function for the whole model or each stratum in case of stratified models. The distribution of the random effect can be plotted using the Gram–Charlier approximation. All other results appear in a text file.

3.2. *Input parameters*

In this section, some of the input parameters of the R interface will be described. Given the wide range of possibilities in the Survival Kit, only a few of its most crucial features are going to be presented.

The general command is:

SKit4R (**program** = "", **discrete** = FALSE, **inputData** = "", **inputPedigree** = "", **time** = "life", **idRec** = "idnum", **censsName** = "cens", **censsValue** = 0, **truncate** = "", **effects** = c("","",""), effectType = c("","","") **pedigreeEffect** = "", **timeDepEffects** = c("","",""),"", **timeDepEffectsType** = c("","","") **classEffects** = c("","",""), **outputEffects** = c("","",""), **strata** = "", **origin** = "", **strataSort** = 0, **ite_quasi** = −1, **model** = c("","",""), **std_error** = TRUE, **random** = "", correlation = c("","","",""), **test** = c("","",""), **baseline** = FALSE, **kaplan** = FALSE, **moments** = FALSE, **survivalOptions** = "", **residual** = "", **graphics** = TRUE)

Out of these function arguments the *program*, *inputData* and *model* are compulsory. The arguments *time*, *censsName*, *censsValue* and *outputEffects* can be skipped if the default settings are used. The rest of the arguments are describing the data structure (e.g., *classEffect*) or trigger optional features (e.g., *test*).

- *program*: Indicates which compiled executable file should be used ("cox" or "weibull"). There is the possibility to bypass the data preparation step and run only the cox or weibull programs by stating "only_cox" or "only_weibull" in case of multiple analyses on the same data set. This is a compulsory parameter.
- *discrete*: The TRUE value indicates that the time scale is expressed in a few discrete units, therefore the grouped data model of [8] is used. This is available with the *weibull* executable file (although it is not a Weibull model!).
- *inputData*: Name of the input data file. This is a compulsory parameter.
- *inputPedigree*: Name of the pedigree file from which a relationship matrix [12] will be constructed.
- *time*: This could be a single variable name holding the name of an already computed survival time, or a set of 3 variables holding the name of the survival time, the beginning and the end of the observation given as dates. In the latter case the total survival time is computed within the program. Only integer time variables are allowed.
- *idRec*: Name of the variable holding the unique identification number for each record.

- *censName*: Name of the variable holding the censoring codes for each record.
- *censValue*: A number that identifies the censored observations in *censName*, everything else is considered as an uncensored record The default value denoting the censored records is 0 in the R interface, but also can be freely specified by the user.
- *truncate*: In case there are left truncated records in the data set, this variable holds the time of the truncation point, which could be either an integer number or a date. If the variable is coded 0 in the data set, the record is treated as not truncated, independently from data type.
- *effects*: A vector holding the names of all effects as they appear in *inputData*, including the names of *time*, *idRec* and *censName* variables.
- *effectType*: This parameter specifies the types of input variables, as required by the Survival Kit. The statement could be omitted when only integer values are used the whole input data set. Otherwise a character vector is expected. The values are "class" for integer, "continuous" for real, "date6" for date in 6 digit format (*ddmmyy*) and "date8" for date in 8 digit format (*ddmmyyyy*).
- *pedigreeEffect*: This specifies the name of the variable which is linked to the relationship matrix (filename in *inputPedigree*).
- *timeDepEffects*: List of time dependent variables.
- *timeDepEffectsType*: The types of time dependent effects should be specified in the Survival Kit, similarly to the *effects* statement. This parameter behaves in the same way as *effectType* above.
- *classEffects*: List of variables to be treated as discontinuous (class) covariates. Everything else stated in *effects* is treated as a continuous covariate.
- *outputEffects*: List of effects to be included to the internal recoded file. In case of large datasets a sizeable amount of space and memory may be saved if only the variables needed for further analysis are included here. If the keyword is omitted, all effects are included into the output data set by default.
- *strata*: Variable name holding the (possibly time dependent) stratification variable in case of stratified Cox, Weibull or piecewise Weibull models. Any number of strata levels is allowed.
- *origin*: Specifies the time points when the hazard should be set to zero in case of piecewise Weibull models. It is also possible to use the same variable name here as in *strata* which resets the hazard at the beginning of each stratum (for example, the beginning of each parity, in domestic animals).
- *strataSort*: In all stratified models the internal recoded files should be sorted according to *strata*. This is largely an automated process, but requires the user to specify the column number of the stratification variable in the recoded file.
- *ite_quasi*: Specifies the number of quasi Newton Raphson iterations (in *weibull*) before switching to full Newton algorithm. If the value 0 is specified, the program starts with the full Newton algorithm. This parameter is not available with the *cox* program.
- *model*: A vector holding the names of covariates from *outputEffects*, to be included in the evaluated model. This is a compulsory parameter.
- *std_error*: If set to TRUE, the asymptotic standard errors of estimates are computed for estimates of all effects.
- *random*: Specifies the name(s) of the random effect(s). Note that these names should also appear in the *model* part. The distribution of the random effects can be specified by the *loggamma*, *normal* or *multinormal* keywords. The variances of the random effects might be known or estimated adding the *estimate* keyword. In this case, the mode of the (approximate) posterior density of the variance is assumed to be the best value for the variance. The mean, standard deviation and the skewness of the distribution could be estimated as well using the *moments* keyword of Survival Kit in the *random* statement.
- *correlation*: Holds the names of two correlated random effects and the value of the correlation coefficient, or the command to *estimate* its value and a starting value for the estimate.
- *test*: If specified the full model is compared to various submodels using likelihood ratio tests to find out the significance of each effect.
- *baseline*, *kaplan*: Setting these to TRUE computes the baseline hazard function and/or the Kaplan–Meier estimate in the *cox* program and visualizes the outcome via the R graphics window.
- *moments*: If set to TRUE, the moments of the approximate posterior density of the variance parameter are computed and used to plot it using the Gram–Charlier approximation (the *moments* should also be specified also in the *random* statement).
- *survivalOptions*: If specified, information about the survivor function of individuals with specific covariate values is produced. It is a very useful tool to relate the estimated regression coefficients to a more conventional scale like median survival time or probability of survival to certain age (only with the *cox* program).
- *residual*: If set to TRUE, it computes the generalized residuals according to [18] for each initial record (only with the *cox* program).

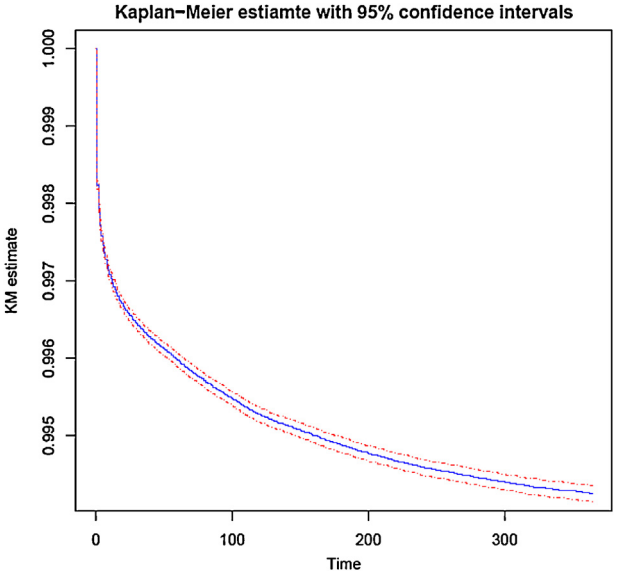

Fig. 1 – Kaplan–Meier estimate.

- *graphics*: If set to TRUE the graphical capabilities of R are used to visualize the results.

4. Applications

4.1. *Infant mortality in Austria*

With this example, we intend to show the use of the Survival Kit with its R interface on data from 2.060.979 records of singleton live births between 1984 and 2008 in Austria. All data were extracted from birth certificates provided by Statistics Austria [19]. The variable of interest was the survival of newborns up to 1 year of age.

```
SKit4R(nrecmax = 2100000,
  inputData = "main.dat",
  effects = c("idnum","life","cens",
    "byear","agem","pregl","gender",),
  effectType = c("class","class","class""class",
    "continuous","class","class"),
  classEffects = c("byear","pregl","gender"),
  ### parameters for the model
  program = "cox",
  model = c("byear","pregl","gender"),
  baseline = TRUE,
  kaplan = TRUE)
```

The input effects were the id number (idnum), length of life (life), censoring code (cens), year of birth (byear), age of the mother (agem), her pregnancy length in weeks (pregl), and the gender of the newborn (gender). From these, the age of the mother was expressed with decimal numbers, therefore the "continuous" specification in the *effects* line (i.e. real values). As it was considered as a continuous effect, it was not specified in the *classEffects* function argument.

Almost all input parameters are described in the section above. Due to the large amount of data, some pre-defined arrays in the Fortran code would not be sufficient, so an additional parameter was used to change the needed size of computer memory accordingly through the *nrecmax* keyword.

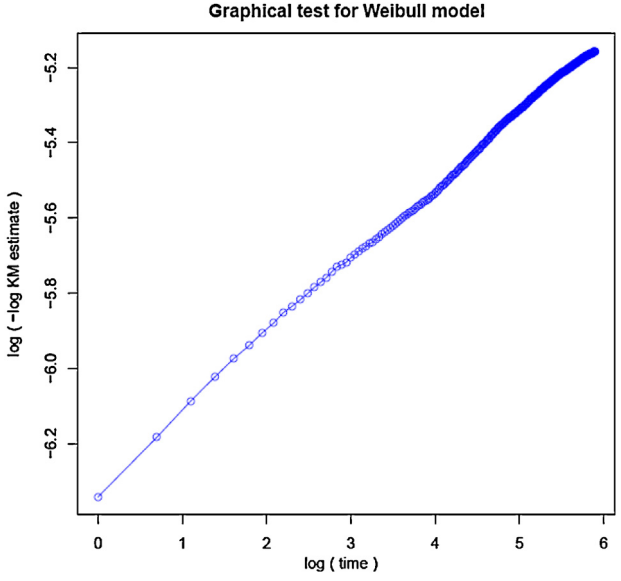

Fig. 2 – Graphical test of the Weibull hazard assumption (a straight line is expected if the Weibull model is adequate).

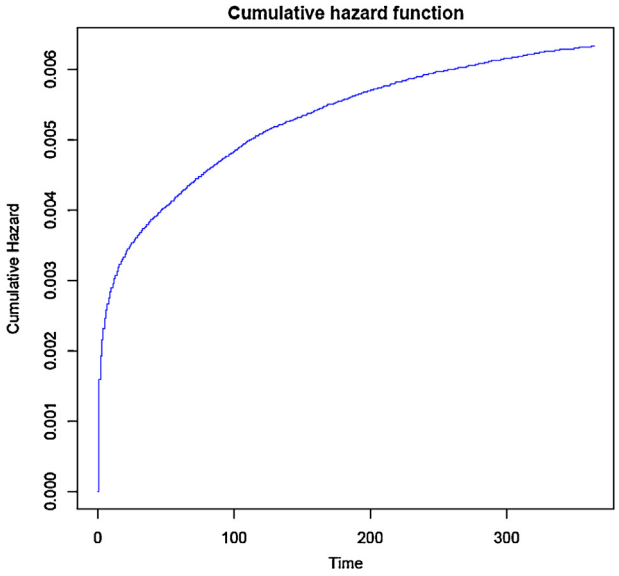

Fig. 3 – Cumulative hazard function.

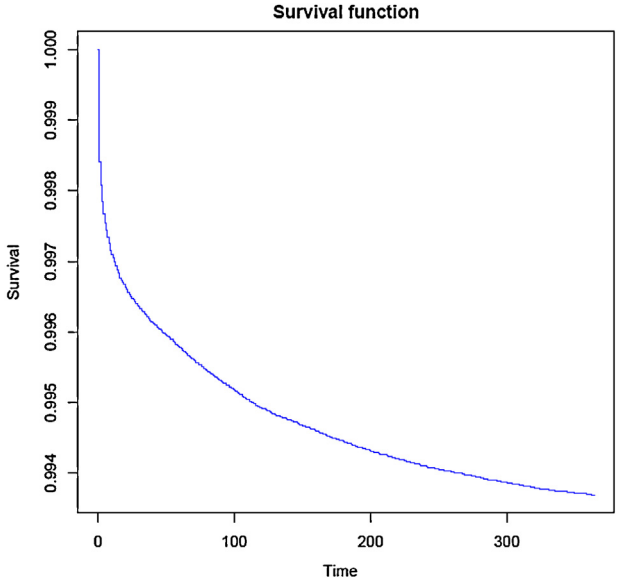

Fig. 4 – Survival function.

For simplicity and illustration, only the effects of the birth year, pregnancy length and gender were calculated. The output of the Survival Kit was automatically saved onto a text file. The relevant parts of this file are shown below. Also, as the *baseline* and *kaplan* options were also specified, a set of graphs was produced by R (Figs. 1–4). Fig. 2 shows a graphical test, where a straight line is expected if the Weibull model is adequate. The calculation of the Kaplan–Meier estimate required for this graph is included in the Cox program. If the sole purpose of the run is to compute the Kaplan–Meier estimate, it is possible to delete all effects from the model statement to increase computational speed, as the values of this estimate are not altered by model covariates.

COVARIATE :		ESTIMATE	STANDARD ERROR	CHI2	PROB >CHI2	RISK RATIO	UNCENSORED FAILURES
1 byear	(DISCRETE)						
1984		0.0000	*	*	*	1.000	902
1985		0.0075	0.0475	0.02	0.8750	1.008	869
1986		-0.0511	0.0486	1.11	0.2926	0.950	799
1987		-0.1152	0.0496	5.40	0.0201	0.891	741
1988		-0.3508	0.0523	44.91	0.0000	0.704	613
1989		-0.2317	0.0505	21.07	0.0000	0.793	695
1990		-0.3045	0.0518	34.59	0.0000	0.737	636
1991		-0.4026	0.0519	60.12	0.0000	0.669	630
1992		-0.4001	0.0513	60.88	0.0000	0.670	658
1993		-0.6009	0.0547	120.64	0.0000	0.548	531
1994		-0.6823	0.0561	148.15	0.0000	0.505	492
1995		-0.8697	0.0593	215.07	0.0000	0.419	416
1996		-0.9415	0.0600	246.16	0.0000	0.390	402
1997		-1.0764	0.0634	287.92	0.0000	0.341	344
1998		-1.0524	0.0636	273.55	0.0000	0.349	341
1999		-1.1355	0.0660	296.12	0.0000	0.321	309
2000		-1.1323	0.0642	310.78	0.0000	0.322	333
2001		-1.1636	0.0671	301.00	0.0000	0.312	296
2002		-1.3600	0.0706	371.32	0.0000	0.257	259
2003		-1.2682	0.0691	336.46	0.0000	0.281	273
2004		-1.2219	0.0672	330.47	0.0000	0.295	294
2005		-1.2911	0.0692	347.79	0.0000	0.275	272
2006		-1.4981	0.0734	416.08	0.0000	0.224	234
2007		-1.4862	0.0747	395.67	0.0000	0.226	224
2008		-1.3563	0.0776	305.46	0.0000	0.258	204
2 pregl	(CONTINUOUS)						
1		-0.4007	0.1584E-02	63968.66	0.0000	*	11767
3 gender	(DISCRETE)						
1		0.0000	*	*	*	1.000	6699
2		-0.1727	0.0186	85.95	0.0000	0.841	5068

According to the results the risk of death until one year of age has been steadily decreasing from the 80s. The highly significant 16% lower risk of death for female newborns (coded as 2) is apparent. The reference class in these cases was automatically set to the class with highest number of uncensored observations (default), but it could be set to any other class. As for the continuous effect, no risk ratio was calculated by the program, because it depends on the width of the interval within the continuous effect. It can be computed manually: e.g. the risk of death associated with a *pregl* of 36 weeks compared to a *pregl* of 40 weeks is increased by $\exp(-0.40 \times (36–40)) = 4.95$.

4.2. *Simulation of correlated random effects*

In a second example, we demonstrate the usage of the Survival Kit when the model includes two correlated random effects. The objective is to simultaneously estimate their variances and their correlation coefficient of two correlated random effects. For this purpose we used two simulated datasets with

Table 1 – Mean and standard deviation (s) across replicates of estimated variances and correlation for different designs and value of true correlation.

		σ_1^2 (true) = 0.3		σ_2^2 (true) = 0.3		ρ (true) = −0.2
		Without ρ	With ρ	Without ρ	With ρ	With ρ
50 levels	$\bar{x}$	0.285	0.285	0.275	0.275	−0.179
	s	0.061	0.060	0.060	0.060	0.150
100 levels	$\bar{x}$	0.279	0.282	0.274	0.277	−0.190
	s	0.040	0.041	0.044	0.044	0.107
		σ_1^2 (true) = 0.3		σ_2^2 (true) = 0.3		ρ (true) = −0.6
		Without ρ	With ρ	Without ρ	With ρ	With ρ
50 levels	$\bar{x}$	0.268	0.284	0.257	0.274	−0.592
	s	0.058	0.060	0.060	0.062	0.101
100 levels	$\bar{x}$	0.262	0.282	0.256	0.277	−0.597
	s	0.039	0.041	0.042	0.044	0.070
		σ_1^2 (true) = 0.3		σ_2^2 (true) = 0.3		ρ (true) = 0.6
		Without ρ	With ρ	Without ρ	With ρ	with ρ
50 levels	$\bar{x}$	0.313	0.285	0.313	0.284	0.622
	s	0.065	0.060	0.073	0.068	0.115
100 levels	$\bar{x}$	0.307	0.283	0.307	0.282	0.615
	s	0.043	0.041	0.052	0.049	0.082

$\bar{x}$, mean of the 200 replicates; s, standard deviation of the 200 replicates; ρ, correlation coefficient between the random effects

different levels of correlation. A single fixed effect with two levels was assumed, and two random effects with 50 levels in set1 and 100 levels in set2. One hundred records were associated with each level of the first random effect in both cases, with a final size of 5000 records for set1 and 10,000 records for set2. The two effects are cross-classified: the first observation of level 1 of random effect 1 is associated with level 1 of random effect 2, the second observation is associated with the second level of random effect 2, etc. The true values of correlation coefficients were either −0.2, −0.6 or 0.6. Each model was evaluated assuming no correlation or estimating the correlation coefficient during the run, with a total of 12 models. Each model was run 200 times. The mean and standard deviation were computed for the random effects variances and the correlation coefficient were computed over the 200 replicates. The R interface for the computation is:

```
SKit4R(inputData = "simData.txt",
  effects = c("idnum","life","cens","fixed","rnd1","rnd2"),
  classEffects = c("fixed","rnd1","rnd2"),
  program = "cox",
  title = "Correlated random effects",
  model = c("fixed","rnd1","rnd2"),
  random = "rnd1 estimate moments normal 0.5
      rnd2 estimate moments normal 0.5",
  correlation = c("rnd1","rnd2","estimate 0.5")
  moments = TRUE)
```

The main difference with the previous example is the occurrence of the *random* and *correlation* statements. This is to specify the names of the random effects (*rnd1* and *rnd2*) and the fact that we want to estimate the variances and the correlation coefficient using the *estimate* keyword. The 0.5 is a starting value for all computations. The starting value used is not so important when estimating the results, but setting it to a value close to the actual solution (using prior knowledge or literature results) might decrease computing time. The results from all runs are summarized in Table 1. They show that correlation was relatively accurately estimated whatever the true value between −0.6 and 0.6. Variances were somewhat underestimated (respectively overestimated) when the correlated nature of the random effects was ignored and the true correlation was negative (respectively, positive). The differences are small, but they may be much larger when the two random effects are not as perfectly cross-classified as in this simulated situation.

When using the *moments* keyword, the mean, standard deviation and skewness of the approximate marginal posterior density of the variance parameters are computed. This can be shown graphically plotting the results from the Gram–Charlier approximation. Fig. 5 shows the comparison of posterior distributions plotted in MS Excel with 100

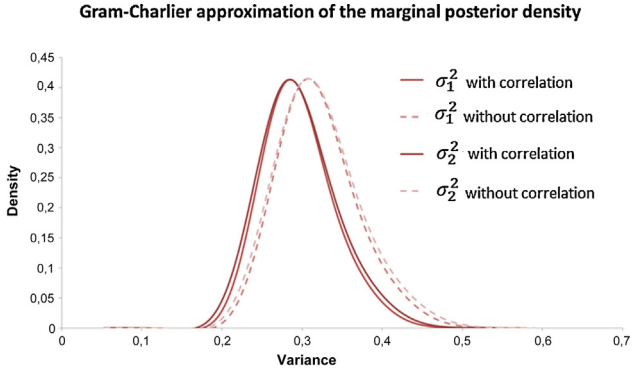

Fig. 5 – Gram–Charlier approximations for the simulated example in Section 4.2 with and without accounting for correlation of 0.6 between random effects.

levels for both random effects with and without accounting for correlation of 0.6 between them.

5. Availability

The Survival Kit can be freely used (including for routine genetic evaluations) provided its use is being credited. Use it at your own risk. The source code, compiled executable files, manual and support programs can be found at: http://www.nas.boku.ac.at/nuwi-survivalkit.html.

Acknowledgements

The authors would like to thank to Thomas Waldhör and Harald Heinzl for their contribution in obtaining and analyzing the infant mortality dataset. The financial support by project number P20552-B17 of the Austrian Science Fund is acknowledged. Significant proportion of this work was carried out during several research stays of the first author at INRA in Jouy-en-Josas, France.

REFERENCES

[1] D.R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society: Series B 34 (1972) 187–220.
[2] J.D. Kalbfleisch, R.L. Prentice, The Statistical Analysis of Failure Time Data, John Wiley and Sons, NY, USA, 1980.
[3] J. Vaupel, K.G. Manton, E. Stallard, The impact of heterogeneity in individual frailty and the dynamics of mortality, Demography 16 (1979) 439–454.
[4] C. Legrand, V. Ducrocq, P. Janssen, R. Sylvester, L. Duchateau, A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model, Statistics in Medicine 24 (2005) 3789–3804.
[5] V. Rondeau, J.R. Gonzalez, Frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation, Computer Methods and Programs in Biomedicine 80 (2005) 154–164.
[6] V. Ducrocq, G. Casella, A Bayesian analysis of mixed survival models, Genetics Selection Evolution 28 (1996) 509–529.
[7] V. Ducrocq, An improved model for the French genetic evaluation of dairy bulls on length of productive life of their daughters, Animal Science 80 (2005) 249–256.
[8] R. Prentice, L. Gloeckler, Regression analysis of grouped survival data with application to breast cancer data, Biometrics 34 (1978) 57–67.
[9] J.D. Kalbfleisch, R.L. Prentice, Marginal likelihoods based on Cox's regression and life model, Biometrika 60 (1973) 267–278.
[10] V. Ducrocq, Extension of survival analysis to discrete measures of longevity, in: Fourth International Workshop on Genetic Improvement of Functional Traits in Cattle: Longevity, Jouy-en-Josas, May 9–11, 1999, Interbull Bulletin, 21, 1999, pp. 41–47.
[11] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Mathematical Programming 45 (1989) 503–528.
[12] R.L. Quaas, Computing the diagonal elements and inverse of a large numerator relationship matrix, Biometrics 32 (1976) 949–953.
[13] V. Ducrocq, J. Sölkner, "The Survival Kit", a FORTRAN package for the analysis of survival data, in: Proc. 5th World Cong. Genet. Appl. Livest. Prod., 22, 1994, pp. 51–52.
[14] V. Ducrocq, J. Sölkner, The Survival Kit – V3.0, a package for large analyses of survival data, in: Proc. 6th World Cong. on Genet. Appl. Livest. Prod., January 11–16, University of New-England, Armidale, Australia, 27, 1998, pp. 447–450.
[15] V. Ducrocq, J. Sölkner, G. Mészáros, Survival Kit v6 – a software package for survival analysis, in: 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, 2010.
[16] L. Tierney, J.B. Kadane, Accurate approximations for posterior moments and marginal densities, Journal of the American Statistical Association 81 (1986) 82–86.
[17] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2009, ISBN: 3-900051-07-0, URL http://www.R-project.org
[18] D. Cox, E.J. Snell, A general definition of residuals, Journal of the Royal Statistical Society: Series B 30 (1966) 248–275.
[19] Statistics Austria: Vital statistics, http://www.statistik.at/web_en/statistics/population/births/index.html